图书情报工作
第65卷 第13期 2021年7月

文本增强与预训练语言模型在网络问政留言分类中的集成对比研究*

■ 施国良 陈宇奇
河海大学商学院 南京 211100

摘 要：[目的/意义]政府网络问政平台是政府部门知晓民意的重要途径之一，为提高问政留言分类的精度以及处理留言数据质量差、数量少等问题，对比多种基于BERT改进模型与文本增强技术结合的分类效果并探究其差异原因。[方法/过程]设计网络问政留言分类集成对比模型，文本增强方面采用EDA技术与SimBERT文本增强技术进行对比实验，文本分类模型方面则采用多种基于BERT改进的预训练语言模型(如ALBERT、RoBERTa)进行对比实验。[结果/结论]实验结果表明，基于RoBERTa与SimBERT文本增强的文本分类模型效果最佳，在测试集上的F1值高达92.05%，相比于未进行文本增强的BERT-base模型高出2.89%。同时，SimBERT文本增强后F1值相比未增强前平均提高0.61%。实验证明了基于RoBERTa与SimBERT文本增强模型能够有效提升多类别文本分类的效果，在解决同类问题时具有较强可借鉴性。

关键词：问政平台 文本分类 文本增强 BERT模型
分类号：G254
DOI：10.13266/j.issn.0252-3116.2021.13.010

1 引言

网络问政是人民群众参与政策决策、维护自身权益的一种新兴民主参与方式[1]。秉承"群众的事无小事"的以人为本原则，近年来，各地政府机构陆续推出了诸如"市长信箱""有话请您说""民意留言簿"等多种形式的网络问政平台。网络问政的兴起可以帮助政府部门更加方便、快捷以及真实地了解百姓们的意见和诉求，大大提升了政府部门的办事效率以及群众满意度。然而，随着信息时代的来临，网络信息量呈现几何式增长，政府网络问政留言亦是如此，人工分类处理手段已远远跟不上数据增长的速度。因此，将自然语言处理(Natural Language Processing，NLP)技术融入"智慧政务"体系具有重大意义。

一般认为，政府网络问政留言分类是提取群众留言中有价值信息的先导工作，以往的留言文本分类方法多是基于人工筛选，往往需要耗费大量人力物力。在引入自然语言处理技术后，传统的文本分类模型往往需要根据不同的任务重新训练词向量来抽取特征，模型效果的好坏和词向量训练语料质量息息相关。与此同时，问政留言数据可能存在较多的无效数据，数据质量的低下会一定程度上影响分类器的效果。

针对以上问题，本文通过对多种BERT(Bidirectional Encoder Representations from Transformers)系列文本分类模型以及不同的文本增强算法模型进行集成对比研究，提出了基于RoBERTa与SimBERT文本增强的政府网络问政留言文本分类模型。期望在以下方面作出贡献：①将预训练语言模型技术用于网络问政平台留言分类任务中，通过多头注意力机制及双向Transformer网络结构缓解传统分类器无法有效解决"一词多义"的矛盾；②针对留言文本的特殊性，采用文本增强模型进行集成对比研究，找出最优组合，一定程度上解决留言文本的数据质量问题，以提高政府部门工作效率；③从模型构造角度分析实验模型表现产生差异的原因，从而更好地为其他领域文本分类任务以及传统自然语言处理下游任务中的文本处理实践提供借鉴和指导。

* 本文系中央高校基本业务费项目"基于图数据库的水利知识图谱关键技术研究"(项目编号:B200207036)研究成果之一。
作者简介：施国良(ORCID:0000-0001-7585-640X)，副教授，博士，E-mail:shigl@hhu.edu.cn；陈宇奇(ORCID:0000-0001-5755-5208)，硕士研究生。
收稿日期:2020-10-07 修回日期:2021-03-19 本文起止页码:96-107 本文责任编辑:杜杏叶

2 研究现状

2.1 文本表示方法研究

文本分类前需要对文本进行建模,抽取其特征并加以表示,选择合适的文本数据特征可以有效提升分类模型的效果。文本特征表示主要分为向量空间文本特征表示、预训练词向量文本特征表示以及预训练语言模型的文本特征表示方法。向量空间模型采用 TF-IDF 算法,并依据词频赋予权重,不过,由于短文本特征存在稀疏性,会造成传统向量空间过于稀疏,进而影响分类结果。基于预训练词向量的文本特征表示则可以有效解决向量矩阵稀疏性问题,马思丹等[2]在训练词向量时将文本关键词分为重叠部分和非重叠部分,并采用参数化线性加权方式计算两部分的相似度,提出了加权 Word2vec 的文本分类方法,效果明显优于传统向量空间算法;程婧等[3]指出训练词向量时常因缺少低频词样本而导致低频词无法进行有效更新,提出可通过与低频词相似的高频词来训练指导低频词更新迭代,从而达到优化词向量的目的,且基于 Word2vec 和 GloVe 得到预训练词向量并用于下游任务,效果显著。

近年来基于预训练语言模型的相关研究表明,无监督的预训练语义表示模型可以有效解决上述两种文本特征表示方法的不足之处。预训练指的是在无标签的文本数据上,以预测句子中的下一个词为目标进行模型训练,从而学习到不同单词的上下文表示关系。M E. Peters 等[4]在 2018 年的 NAACL 会议中提出的 ELMo(Embeddings from Language Models)模型通过双向长短时记忆网络(LongShort-Term Memory, LSTM)结构构建动态词向量,缓解了词的多义性问题,从此开启了 NLP 任务中的预训练语言模型时代。在此基础上,Google 的 J. Devlin 等[5]于同年 10 月提出了 BERT 模型,BERT 模型采用语义表征能力更强且融入了自注意力机制的 Transformers 模型[6]代替 ELMo 中的 LSTM 结构,同时利用海量公开语料进行训练,很大程度上提升了预训练语言模型的动态词向量表征能力。如果说 ELMo 模型开启了预训练语言模型的时代,那么 BERT 模型则是通过注意力机制和海量训练语料将预训练语言模型推向高潮的代表。

2.2 文本分类算法研究

文本特征抽取可以有效将文本转化为特征向量以支持后续任务,分类算法则是在此基础上区分短文本的特征,将其划分至正确的类别中。随着研究的深入,越来越多的学者将机器学习算法应用到文本分类任务中并取得了不错的效果。陈燕方等[7]构建了在线商品可信度因素指标,并在此基础上将指标体系融入 SVM 分类器中,提出了 DDAG-SVM 在线商品评论可信度分类模型。余本功等[8]采用 SVM 与随机森林的多通道的建模方式,提出了 nLD-SVM-RF 的短文本分类算法,提高了模型的泛化性能。

与此同时,随着数据量增加以及计算机性能的提升,深度学习算法在文本分类领域的优势也逐渐显现了出来。韩栋等[9]赋予主题句较高的权重并将其融入字符级卷积神经网络(Convolutional Neural Networks, CNN)中进行文本分类研究;杨云龙等[10]为解决单一循环神经网络(Recurrent Neural Network, RNN)无法长时间记忆问题,提出了融合胶囊特征的门控循环单元(GRU)的文本情感分析模型 G-Caps,有效提高了中文情感分析的效果。在将预训练语言模型作为基础分类器的研究方面,赵旸等[11]运用中文医学预训练模型(BERT-Re-Pretraining-Med-Chi)进行文献分类研究;吴俊等[12]运用 BERT 模型进行文本向量化后接入 BiLSTM-CRF 模型进行中文专业术语的命名实体识别(Named Entity Recognition, NER)研究,相比于传统预训练词向量效果显著提升;廖胜兰等[13]则是将 BERT 模型作为"教师模型"进行模型蒸馏,以提升 Text-CNN 等轻量分类器的分类效果,从效果和量级双方面对分类算法进行了优化。

2.3 文本增强算法研究

政府网络问政留言属于短文本的一种,相较于长文本,一般篇幅较短,并且具有较强的随意性与不规范性,尤其是留言文本,其中充斥着大量的网络用语、口头语以及简称,导致文本噪音较大,质量合格的文本数量有限。文本数据增强技术可以一定程度上缓解上述问题,W. Jason 等[14]于 2019 年总结提出了系统性文本增强策略(Easy Data Augmentation, EDA),主要通过词语层面的变化生成新的句子来达到文本增强的效果;俞畅等[15]则是以 RNN 作为生成网络,CNN 作为判别网络,提出基于生成对抗网络(seqGAN)的电力用户意图文本的生成模型,并采用 BLEU 算法验证了生成文本的有效性。

2.4 研究不足与总结

文本表示方法与分类模型方面,尽管预训练词向量的文本表示方法通过词嵌入将不同的词(token)映射成单一向量有效缓解了传统向量空间表示法的不足,但仍存在以下问题:①训练词向量之前需要进行分词,分词词库的不精确导致无法有效识别未登录词,这

会影响向量表示准确度。②不同情境下的文本表示任务(医学背景或法律背景等)运用同一套预训练词向量无法达到最佳效果,而根据自身任务训练词向量需要大量训练语料和训练设备的支撑,实践可行性较低。③预训练词向量可以解决“一义多词”问题,却不能解决“一词多义”问题。因而,本文实验选用预训练语言模型进行文本特征表示与分类。与此同时,相比于体量庞大的 BERT 模型,本研究在此基础上进行改进,采用基于 BERT 改进的 RoBERTa[16] 与 ALBERT[17] 预训练语言模型进行网络问政留言的分类,并尝试后接神经网络作为分类器的方法提升模型的效果。

文本数据增强方面,EDA 文本增强技术多基于规则变化,生成的文本的向量特征表示可能与原文差别不大从而导致训练样本重复无效;与此同时,通过生成对抗网络生成的文本都是随机无规则的领域文本,文本的类别标签也需通过模型的判别网络预测给出,必然会存在一定的误差从而影响模型的训练。因而针对上述问题,本文研究采用了 SimBERT[18] 文本增强技术,SimBERT 主要以 BERT 模型为基础利用有监督相似文本对训练而成,可以针对特定的句子生成其相似句,标签则使用原始数据标签,同时解决了生成文本的一词多义问题与标签不精确问题。

3　模型设计与整体框架

为提高政府网络问政留言分类准确性以推进“智慧政务”服务体系构建,本文设计了网络问政留言分类集成对比模型,以近年 NLP 领域较为流行的 BERT 预训练语言模型及其改进模型作为文本表示模型并结合 EDA 及 SimBERT 文本增强算法模型完成留言文本分类任务。

实验模型总体设计框架见图 1。文本分类模型方面,本文选取预训练语言模型 BERT-base 作为基线模型,通过字符级粒度嵌入文本特征向量完成网络问政留言分类任务,并选取 BERT 及其附加网络模型、基于 BERT 改进的预训练语言模型 ALBERT 与 RoBERTa 进行对比实验。与此同时,为缓解网络问政平台留言口语化严重、数据质量低等问题,选取了基于规则增强文本的 EDA 文本增强算法与融合了自然语言生成(Natural Language Generation,NLG)与自然语言理解(Natural Language Understanding,NLU)的相似句生成模型 SimBERT 作为文本增强模型。旨在通过模型间的集成对比实验,同时结合政府网络问政留言数据自身的特点,设计出解决同类问题的最优模型组合,并在此基础上从模型构建原理的角度分析模型表现差异的原因。

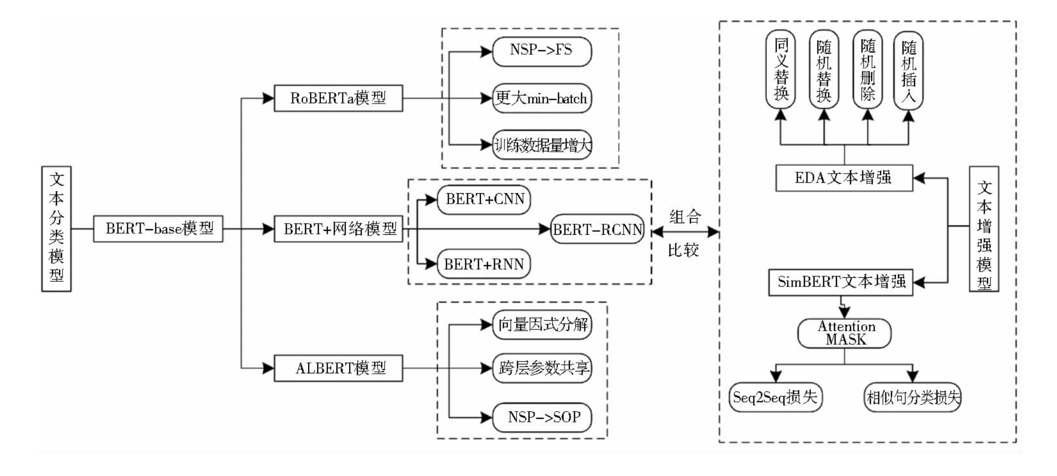


图 1　网络问政留言分类集成对比模型

3.1　文本分类模型选择与设计

选择 BERT 及其改进预训练语言模型生成文本字嵌入向量解决文本分类任务主要出于以下考虑:①预训练语言模型可以不依赖于传统的人工特征抽取,以端到端的形式完成文本的向量化表示并加以微调,进而应用于其下游任务中;②基于双向 Transformer 结构及注意力机制的 BERT 模型在文本特征抽取时可以有效解决“一词多义”等问题;③中文文本存在字和词两种不同的划分粒度,传统的预训练词向量多在分词后进行训练,分词的过程中难免出现误差,而 BERT 可以基于字符级粒度进行字嵌入,在中文任务中一般有更好的表现。

BERT 全称为多层双向变换器编码器[5],是 Google 公司于 2018 年底基于融入自注意力机制的 Transformer[6] 结构提出的预训练语言模型。BERT 的优势在于其强大的词向量泛化能力,不同于传统的 one-hot 编码

以及 word2vec 预训练静态词向量，BERT 通过双向 Transformer 结构动态调整词向量，充分融入词语上下文信息，可以较好地解决一词多义问题。BERT 模型的结构示意图如图 2 所示：

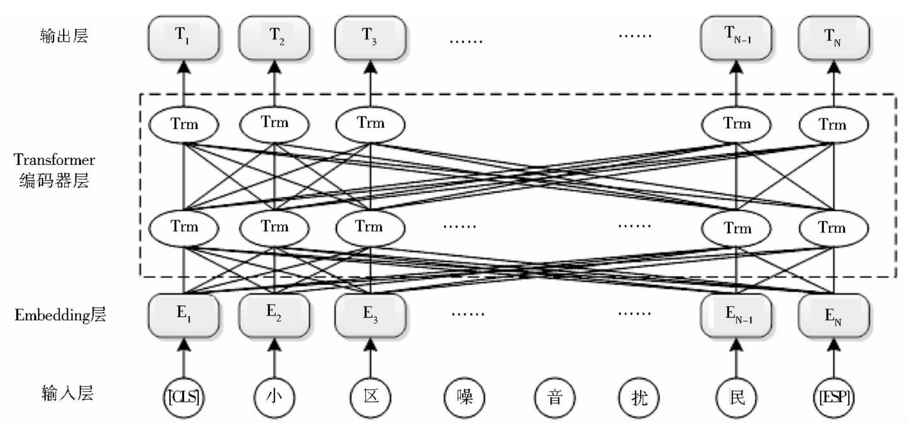


图 2　BERT 模型结构示意

图 2 中 E_1，E_2，…，E_N 表示输入文本的字词级 Embedding 层，而后经过双向 Transformer 编码器得到经过 Self-attention 机制融入上下文信息的输出 T_1，T_2，…，T_N，在原始 BERT 模型中，将文本的输出传入一个 *softmax* 层即可得到最终的分类结果。

BERT 模型内部的基础结构为 Transformer 模型[6]，该模型是基于 Self-attention 机制的 Seq2seq 模型，是典型的 Encoder-Decoder 结构模型，主要是通过 Encoder 层将输入序列编码为固定长度向量，再经过 Decoder 层将固定长度向量解码为任务所需长度的输出序列。Transformer 模型中 Encoder 模块的输入是文本的词嵌入表示，并且融入了位置编码信息 Positition Encoding。模型的核心在于替换传统 RNN 及 CNN 结构的 Self-attention 层，Self-attention 层相当于序列编码层，主要作用在于将某个词与句子中其他部分内容的关系融入该词的词向量中，从而解决一词多义问题。其主要原理及计算步骤如下：

（1）输入的句子文本以字词为单位嵌入成词向量。

（2）将词向量与权重矩阵 W^Q，W^K，W^V 分别相乘，得到与之对应的 *Queries*（*q*），*Keys*（*k*）以及 *Values*（*v*）向量，其中 W^Q，W^K，W^V 的维度分别为 $N*d_k$，$N*d_k$，$N*d_v$，*Queries* 与 *Keys* 向量维度均为 d_k，*Values* 的维度为 d_v。

（3）计算每个向量的 *score*，$score = q \cdot k$，该分数为对句子中某个词进行 encoder 时，该词对句子中其他部分的关注度。

（4）对不同单词对应的 *score*，利用 *softmax* 激活函数将其转化为取值在 0 – 1 内且总和为 1 的数作为权重 w。

（5）经 *softmax* 后得出的权重值与 *v* 相乘，得到加权的评分向量，最后再进行相加求和得出最终的输出 z，$z_i = \sum w_i * v_i$。

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad 公式(1)$$

$$Head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad 公式(2)$$

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_n)W^O \quad 公式(3)$$

上述公式（1）是 Self-attention 机制的计算公式，公式（2）、公式（3）则是 Multi Head Attention 的计算公式，其中 W_i^Q，W_i^K，W_i^V 表示第 i“头”自注意力机制中的权重矩阵，W^O 为全连接层的矩阵，将不同“头”的自注意力机制输出横向拼接之后经过全连接层构造指定维度的最终输出结果矩阵。

与此同时，BERT 模型的主要创新贡献之一在于其独特的预训练方式，BERT 采用遮蔽语言模型（Masked Language Model，MLM）以及下一句预测模型（Next Sentence Prediction，NSP）作为任务进行预训练，可以有效提升模型的深度双向预测能力以及推理能力。BERT 模型的输入向量主要由词向量、段向量以及位置向量三者加权求和组成，句子开头和结尾分别采用［*CLS*］及［*SEP*］标识，句子间也采用［*SEP*］标识进行分割，具体结构见图 3。

MLM 模型主要是以 15% 的概率抹去句子中的一个或几个词，训练模型利用剩余的字词去预测所遮蔽的字词，类似于完形填空任务。这样做的目的是为了在不影响模型理解能力的基础上防止因过多采用［*MASK*］标识导致模型预训练的效果下降。不同于一

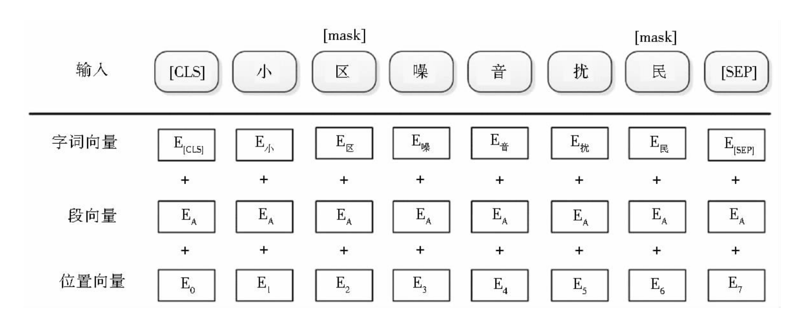


图 3 BERT 模型输入向量组成

般的双向 LSTM 只能训练模型从左至右以及从右至左分别理解左侧上文以及右侧下文的信息，遮蔽语言模型可以双向深度训练 BERT 模型的句间信息理解能力。

NSP 模型主要目的在于训练模型的句子级别间的上下文关系，主要通过输入语料库中的大量句子对 AB，其中句 B 是句 A 下一句的概率为 50%，句 B 是在语料库中随机选择的句子的概率为 50%，两个句子间以［*SEP*］标识隔开，模型则根据这些句子对数据进行二分类预测训练，以此提高自身理解句子级别间关系的能力。

另外，本文实验在此基础上对 BERT 进行改进，将 BERT 输入层的输出向量结果接上其他的文本分类器，此处我们选择循环神经网络 RNN、卷积神经网络 CNN 以及循环卷积神经网络 RCNN 作为后接的文本分类器。Text-RNN 模型是 P. F. Liu 等[19]提出的基于循环神经网络及其变种（LSTM、GRU）的文本分类器，通过双向循环神经网络可以一定程度上融入上下文语义，能够达到较好的分类效果，但模型训练速度较慢。Y. Kim[20]首次将处理图像问题的卷积神经网络应用到文本分类任务中，提出了 Text-CNN 模型，其原理是把句子或词看成词向量矩阵作为模型的输入，经过卷积层和池化层提取句子的重要特征后进行分类。S. W. Lai 等[21]将 Text-RNN 模型与 Text-CNN 模型相结合，提出了 Text-RCNN 模型，模型通过 RNN 的双向循环结构捕捉上下文信息，并通过 CNN 的最大池化层捕捉关键信息，解决了 Text-RNN 模型的偏倚性问题以及 Text-CNN 模型中固定窗口提取特征的弊端。

3.1.2 RoBERTa 模型

2019 年，Facebook 提出了基于 BERT 改进的预训练语言模型 RoBERTa[16]，其在模型层面并没有对 BERT 的结构作出改动，主要是针对模型的预训练方式进行了调整和优化，并在当时众多下游 NLP 任务中达到了 SOTA，相比于原始的 BERT，RoBERTa 主要作了以下改动：

（1）BERT 在 MLM 预训练任务中，以 15% 的概率随机对句子中的 Tokens 进行 MASK，但是采用这种方式进行 mask 时，一旦 Tokens 被选定了，在接下来的整个训练过程中将无法改变，因此这种方式属于静态 Masking 法。RoBERTa 在 MLM 预训练任务中则采用了动态 Masking 的方式，在训练开始前将全部数据复制 10 份，并对 10 份数据分别进行 MASK，这样同样一句话就有 10 种不同的 Masking 方式。RoBERTa 通过对 Masking 方式的改进，在不同的任务中，模型性能平均提高了 0.3%。

（2）RoBERTa 不再采取 BERT 的 NSP 预训练任务，而是通过每次输入多个连续的句子来训练模型捕捉句子间关系的能力，这种预训练方式被称为 FULL SENTENCES，改进后 RoBERTa 在句间关系推断任务上效果得到提升。

（3）RoBERTa 参数设置上采用了更大的 mini-batch。相比于 BERT 的 256 batch size，RoBERTa 将其调整到了 8k，更大的 batch size 需要配合更大的 learning rate，可以在提升模型训练速率的同时提高模型的效果。

（4）RoBERTa 的预训练数据大小是 BERT 的 10 倍。RoBERTa 使用总计 160GB 的文本训练数据，除了 BERT 本身的训练数据外，还包含了诸如 Web 文本语料库（38GB）、Common Crawl News 数据集（76GB）等数据，并在更大的 GPU 集群上训练了更久的时间。

3.1.3 ALBERT 模型

为了进一步提高模型的性能，XLNet[22]和 RoBERTa[16]模型在优化原始 BERT 预训练方式的同时，都加大了训练数据的量和训练的时长。不过，该做法会致使模型参数量过大，虽然当一个模型的参数量逐渐变多时，模型的效果会有所提升，但是当模型复杂程度过高、参数量过大时，模型的效果反而会降低，这种现象成为“Model Degratation”。为解决这一问题，有研究者

利用知识蒸馏(Knowledge Distillation,KD)方法缩减模型参数,其中以DistillBERT[23]以及华为诺亚方舟实验室提出的TinyBERT[24]模型为代表,它们均能达到了缩减模型大小以及减少参数的目的。采用知识蒸馏的方法虽然可以降低模型的量级,提高模型的计算速度,但却会以牺牲模型的性能为代价,以TinyBERT为例,模型大小仅为BERT的13.3%,参数量为BERT的28%,但是在GLUE基准上相比于BERT却下降了3个百分点。

针对上述问题,Google的蓝振忠团队[17]提出了ALBERT模型,该模型通过采用对Embedding层向量进行因数分解以及跨层参数共享等改进措施,成功做到了在模型参数量缩小18倍的情况下,模型的性能反而超过了BERT、XLNet等大规模预训练语言模型。ALBERT相比于BERT主要进行了三方面改进:

(1)Embedding层向量因式分解。BERT中词语的embedding和encoder后最终输出的embedding维度都是768,ALBERT的研究团队则认为词的原始embedding相比于隐藏层的输出embedding所蕴含的信息量要少的多,因此可以减少词语级别embedding的维度。方法主要是通过将one-hot向量映射到低维空间中减少词语级别embedding的维度(E),随后再映射到高维的空间保持隐藏层的encoder输出维度(H)不变。最终将模型的参数量从$O(V*H)$降低到了$O(V*E+E*H)$。

(2)跨层的参数共享。Transformer模型中提出了共享参数的方法,但是只共享全连接层或者Attention层的参数,ALBERT则是将二者相结合,同时共享编码层的全部参数,在大大减少了模型参数量的同时提升了模型的训练速度。虽然模型效果有所下降,但可以用更大规模的数据量进行训练以提升模型的效果。

(3)RoBERTa预训练过程中,用FULL SENTENCES任务代替了NSP任务,而ALBERT则是基于NSP任务做出了改进,提出了句子顺序预测(Sentence-Order Prediction,SOP)预训练任务。相比于NSP任务,SOP任务可以更好地训练模型句间关系推理的能力,其本质依旧是训练一个二分类器,且正样本与NSP任务相同,不过负样本则改为预测两个相邻的句子是否为逆序句子对。

与此同时,ALBERT还做了诸如移除dropout层等改进,真正意义上同时做到了缩小模型大小和提升模型性能的地步,为BERT系列模型真正实现工业界落地打下了坚实的基础。

3.2 文本增强模型选择与设计

文本分类任务中不同场景下对应的数据质量和数量都会有所差距,本文所研究的政府网络问政留言数据也存在着噪声大、合格数据量少等问题。为缓解因训练数据本身问题所导致的分类模型效果下降的问题,文本采用了EDA[14]以及SimBERT[18]文本增强技术进行政府网络问政留言数据增强,并进行了相关对比实验以探究不同文本增强技术的效果及其优缺点。

3.2.1 EDA文本增强

作为系统性文本增强策略的代表,EDA文本增强算法主要采用传统的基于规则层面的文本增强方法,即对句子词语及语法层面做出相应的修改操作,EDA文本增强的主要数据操作及示例如表1所示:

表1 EDA文本增强数据操作

数据操作	具体方法	示例数据
同义词替换	随机抽取句子中的N个单词,对其进行同义词替换	A市魅力之城小区楼下噪音扰民
随机替换	选中句子中的随机两个单词并交换其位置,可重复多次	小区A市魅力之城楼下噪声扰民
随机删除	指定数值概率p(p为参数),句中的每个单词以概率p进行删除	A市之城小区楼下噪声扰民
随机插入	从句子中抽取一个单词,并将其同义词插入句中的随机位置	A市魅力之城小区美丽楼下噪声扰民

3.2.2 SimBERT文本增强

经实验,EDA文本增强技术可以一定程度上提升模型的表现,但基于规则的数据增强方法仍存在不足,为进一步提升文本增强的效果,本文尝试采用SimBERT文本增强模型[18]。SimBERT是以BERT模型为基础,采用了微软提出的UniLM模型[25]训练思想的融生成与检索为一体的生成式语言模型。

UniLM以多层Transformer模型为主体架构,该模型融合了自然语言生成与自然语言理解的功能。UniLM本质是一种统一的预训练语言模型,主要通过特殊的Attention MASK方式将多个相异的语言模型以共同目标进行联合预训练,并且在训练过程中通过单个Transformer模型对不同的语言模型实现参数共享,此类参数共享方式可以让模型能够同时学习并融合不同的文本特征表示,从而达到联合优化的效果。UniLM主要是在Bidirectional LM、Unidirectional LM以及Sequence-to-Sequence LM这三种语言模型上进行联合训练。

SimBERT 则是借鉴了 UniLM 中 Seq2Seq 部分的训练方式，属于有监督训练，训练语料是所收集到的大量相似文本对，主要训练目标是构建能够预测给定句子相似句的 Seq2Seq 部分。在 SimBERT 的训练中，相似句对中的不同句子通过[*SEP*]标识符隔开，并在此基础上运用特殊的 Attention MASK 方式，即在[*SEP*]前半部分句子中的每个 tokens 之间做双向 Attention，后半句的 tokens 间则做单向 Attention 操作，模型可以递归预测后半句，从而具备 NLG 的能力。在此基础上，SimBERT 还在输入时加入了随机[*MASK*]，这样模型在训练的过程中可以做 MLM 任务，MLM 任务可以训练模型的 NLU 能力。

训练过程中，SimBERT 将每个 *batch* 内所有的[*CLS*]句向量拼接形成句向量矩阵 $D \in R^{b*d}$（其中 b 为 *batch_size*，d 为 *hidden_size*），并在 *hidden_size* 维度上做 L_2 正则化后得到正则矩阵 $\tilde{D}$，内积运算后得到最终的相似度矩阵 $\tilde{D}\tilde{D}^T \in R^{b*b}$，其中对角线部分被 MASK 掉。SimBERT 通过相似度矩阵让模型做分类任务，其中负样本即不相似文本，并借助 *softmax* 操作来增加正样本的相似度，同时降低负样本的相似度。最终 SimBERT 的损失函数即为 Seq2Seq 损失与相似句分类器中 *softmax* 层损失相加的联合损失函数，训练方式示意图如图 4 所示：

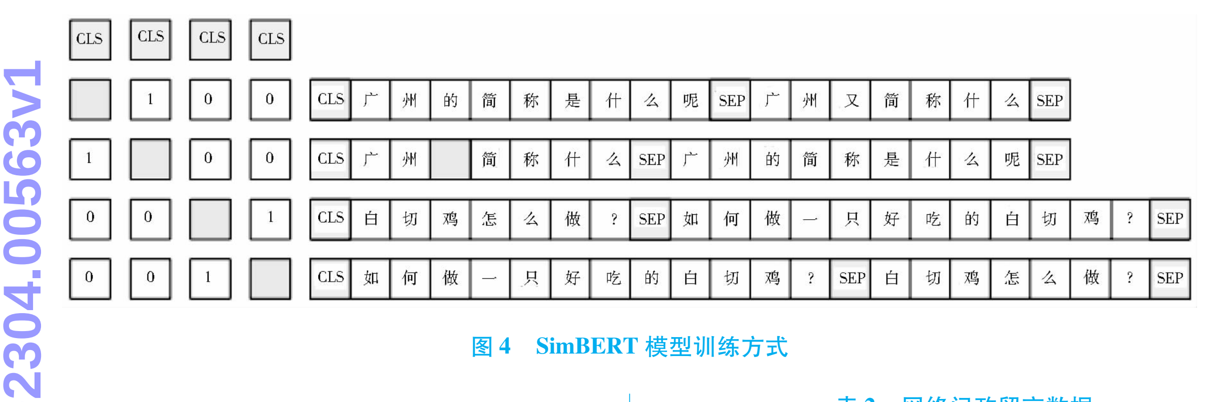


图 4 SimBERT 模型训练方式

4 实验过程与结果分析

4.1 数据来源与预处理

本文的实验数据均来源于我国某省相关网络问政平台 2014 – 2020 年的部分真实留言数据，检索时间为 2020 年 4 月，共计 9 281 条。数据包含城乡建设、劳动和社会保障、教育文体、交通运输等 7 个类别。数据预处理方面，对获取的数据以留言文本长度、重复数据等为标准进行去重，并对留言文本中所包含的市、区、县、镇等地点敏感词汇进行脱敏处理，最后以 8:2 的比例划分训练集和测试集并在训练集中以相同比例划分出验证集，从而得到最终的实验数据集，部分数据展示见表 2，留言类别分布情况见图 5。

表 2 网络问政留言数据

留言文本	留言所属类别
A2 区泰华一村小区第四届非法业委会涉嫌侵占小区业主公共资金	城乡建设
E1 区液压件厂金星村居民区内有很多机械配件加工厂，污染环境	环境保护
A 市交通运输局外十字路口红绿灯不亮，交通事故频发	交通运输
关于尽快建立 F 市民办学校教师社会保险制度的建议	教育文体
《关于开展城乡居民大病保险工作的指导意见》是怎样实施执行的呢	劳动和社会保障
请 M2 县相关部门调查这样品牌经营模式，以及是否涉嫌违法	商贸旅游
I3 县大通湖区无法办理流动人口证明婚育证明和准生证	卫生计生

由网络问政留言数据类别分布情况可知，城乡建设、劳动和社会保障和教育文体类别的留言较多，这三类贯穿了人们日常生活的基本层面。城乡建设中大多数问题反应的是小区物业问题以及周遭生活环境问题等，拥有好的住宿条件及环境是人民群众的安生立命之本，是开展其他社会活动的最根本前提；劳动和社会保障问题则与人们自身的利益紧密相关，解决好人民劳动和社会保障问题也是促进社会公平的重要途经之一；教育文体问题关乎自身及子女的学习发展，也是在满足基本生活需求前提下人们对于自身更好发展的追求。因此，在网络问政留言中人们对上述三类问题关注度最高也是最迫切希望得到解决的问题。如何有效地从大量、繁杂的留言中正确识别出人们关注的问题所属类别是提高政府部门行政办事效率的基础性工作，也是保障人民群众基本利益的重要手段之一。

4.2 实验设置

4.2.1 实验环境设置

本文实验主要在 PyCharm 上运行，实验语言采用 Python 3.7.3，具体的实验环境与软硬件相关配置见表 3。

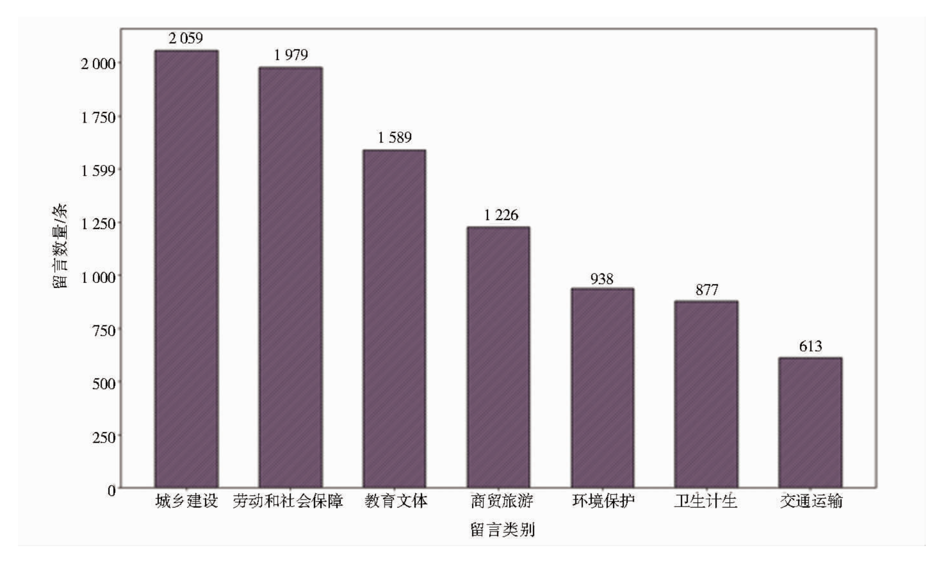


图 5　网络问政留言类别分布



表 3　实验环境配置

实验环境	具体配置
操作系统	Windows10
CPU	Intel Core i7 –9750H
GPU	NVIDIA Tesla T4(16GB)
内存	16G
Python	3.7.3
TensorFlow	1.13.1
Pytorch	1.3.1
Keras	2.3.1

4.2.2　模型参数设置

本文的网络问政留言分类集成对比模型主要分为文本分类模块与文本增强模块,为实现模型在验证集上的最佳效果,具体的参数设置如下。

文本分类模型方面,选取三种不同的中文预训练语言模型,分别是开源的 bert_base_chinese 模型、albert_base_chinese 模型以及 robert_base_chinese 模型。其中 BERT 模型共计 12 层,采用 12 头注意力模式,隐层为 768 维,模型的参数量为 110M;ALBERT 模型共计 12 层,隐层为 128 维,参数量为 12M;RoBERTa 模型共计 6 层,隐层为 384 维,参数量为 200M。BERT 基准模型中,文本长切短补 pad_size 设置为 128,批量训练大小 batch_size 为 32,初始学习率 learning_rate 为 2e-5,采用的优化器为 BertAdam。ALBERT 模型与 RoBERTa 模型的优化器为 AdmLR,learning_rate 设置为 1e-4,其余参数与 BERT 基准模型相同。

文本增强模型方面,EDA 参数中,句中每个词被替换的概率 alpha 设置为 0.3,生成个数 num_aug 设置为 1。SimBERT 参数中,n 设置为 25,k 设置为 1,其中 n 表示用 seq2seq 生成的 n 个相似句,k 表示在生成的相似句中,经过 encoder 计算相似度后返回最相似的 k 个句子。

4.3　实验结果分析与讨论

4.3.1　模型效果评估指标

网络问政留言分类问题属于文本分类问题,为检验对比模型的效果,本文采用精确率(Precision,P)、召回率(Recall,R)以及 F1 值(F-score)作为模型效果衡量指标。精确率(P)、召回率(R)以及 F1 值计算公式如下所示:

$$P=\frac{TP}{TP+FP} \quad \text{公式(4)}$$

$$R=\frac{TP}{TP+FN} \quad \text{公式(5)}$$

$$F1=\frac{2*P*R}{P+R} \quad \text{公式(6)}$$

其中,精确率(P)表示真的正样本预测为正样本(TP)的个数占所有预测为正样本个数($TP+FP$)的比例;召回率(R)表示真的正样本预测为正样本(TP)占真实正样本个数($TP+FN$)的比例;F1 值则是精确率与召回率的调和平均指标,可精确反映出模型多方面效果的好坏。

4.3.2　集成对比模型效果分析

为对比不同文本增强技术以及预训练语言模型在网络问政留言分类任务上的集成对比效果,本文设计了网络问政留言分类集成对比模型,并在训练集上进行训练,通过验证集优化训练结果,最终在测试集上通过上述指标进行模型效果评价。具体实验结果如表 4 及图 6 所示:

表 4 网络问政留言分类模型效果对比

模型	原始数据 F1 值/%	EDA 增强 F1 值/%	SimBERT 增强 F1 值/%
BERT-base	89.16	89.18	89.25
BERT + RNN	88.32	89.54	89.33
BERT + CNN	89.51	89.63	90.29
BERT + RCNN	90.68	91.25	91.52
ALBERT	90.89	91.86	91.05
RoBERTa	91.28	91.93	92.05

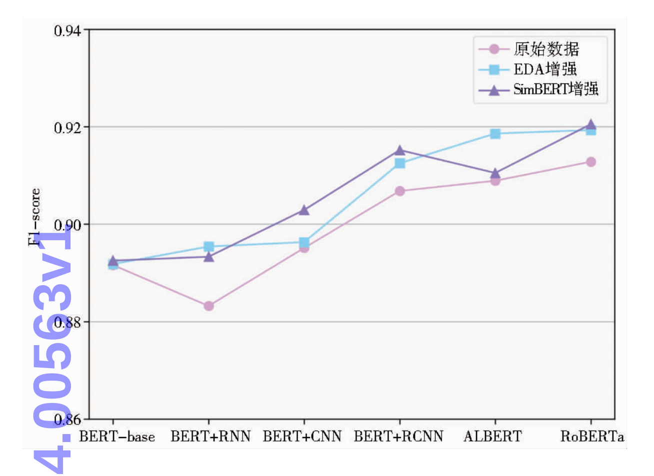


图 6 网络问政留言分类模型效果对比

（1）网络问政留言文本分类效果分析。由上述模型对比结果可知，在不考虑文本增强的前提下，RoBERTa 模型在网络问政留言分类任务上的表现最佳，F1 值达到了 91.28%。ALBERT 模型的 F1 值也高达 90.89%，BERT 系列模型的 F1 值平均为 89.42%，明显略低于 RoBERTa 与 ALBERT 模型的分类效果。

一方面，前四个模型中 BERT-base、BERT + RNN 与 BERT + CNN 的模型效果区别不大，BERT + RCNN 模型的分类性能明显高于前三者，甚至在文本增强后其效果均超过了未进行文本增强的 ALBERT 模型。究其原因，是因为 RCNN 网络结构融合 RNN 与 CNN 网络结构的优点，既可以通过双向循环的 RNN 结构捕捉到句子的上下文信息，又可以通过 CNN 中的最大池化层捕捉到句中的关键信息，可以更加准确地表达句子的语义结构。

另一方面，ALBERT 与 RoBERTa 模型的效果更好主要是因为二者都是基于 BERT 的改进预训练语言模型，相比于 BERT，都采用了更大的训练数据量和更长的训练时间。RoBERTa 主要在 BERT 的预训练任务上进行了改进，将静态 Masking 转化为动态的 Masking 的同时，提出了 FULL-SENTENCES 预训练任务训练模型的句间理解能力。ALBERT 则是在将 NSP 预训练任务改为 SOP 预训练任务时，提出了通过词嵌入层因式分解和跨层参数共享的方式大幅度减少了模型的参数量，从而可以在同等的时间及空间复杂度上可以用更多的数据进行模型训练。在论文[17]中，ALBERT 在绝大多数任务中表现都要好于 RoBERTa，但本文中 RoBERTa 的 F1 值却比 ALBERT 高出 0.39%，除了不同数据背景的影响外，还因为问政留言分类属于标注任务，在标注任务中，RoBERTa 的效果要略优于 ALBERT，与本文实验结果一致。

（2）网络问政留言文本增强效果分析。文本增强方面本文主要采用基于规则的 EDA 文本增强算法与基于相似句生成的 SimBERT 文本增强模型。在进行数据预处理及划分训练集、验证集以及测试集后，通过文本数据增强技术将训练集的数据量扩充至原来的两倍，并重新进行模型训练，与未进行数据增强前的模型进行比较与分析。两类模型下未增强前数据与增强后数据的对比结果如表 5 所示：

表 5 文本增强数据对比展示

序号	未增强数据	EDA 增强后数据	SimBERT 增强后数据	留言类别
1	K3 县人民医院主治医生却开外面药房的药	K3 县治民众医院主治医生却放外面药房的药	K3 县人民医院主治医生开了外面药店的药，怎么办	卫生计生
2	对 K5 县打造洄龙塔的一些建议	对 K5 县营造洄龙塔的些许意见	如何评价 k5 县的洄龙塔	教育文体
3	投诉 G5 县小渡口镇政府暗箱操作招标项目	举报 G5 县小河上镇政府暗箱操作招标计划	投诉 g5 县小渡口镇政府招商招标怎么投诉	城乡建设
4	强烈要求箴言中学退还自主招生中违规收取的 7800 元	强烈要求十诫初中退还自主招生中不合规收取的 7800 元	箴言中学自主招生中违规收费 7800 元怎么办	教育文体
5	反映 A2 区青园花都电梯安全问题	反映 A2 区青园花都楼梯安全原因	a2 区青园花都电梯安全事故如何处理	商贸旅游

表 5 展示了相同的文本内容经过 EDA 和 SimBERT 模型文本增强后的效果，我们可以看到，EDA 文本增强正如其原理一样，仅仅是基于规则进行个别词语的调序和替换，而 SimBERT 文本增强则不是简单的调序，更倾向于以疑问句的方式对原始语句进行改写。

与此同时，由表 4 及图 6 的模型效果对比结果可

见,由于数据量和数据质量的原因,文本增强前后模型的分类效果也有明显区别。对比实验结果可知,经过EDA和SimBERT文本增强后的模型F1值较文本增强前分别平均提高了0.59%和0.61%,证明在数据量有限或数据质量低下的前提下,通过文本增强模型构造训练数据,确实能够一定程度上提升模型的效果。

SimBERT文本增强后的F1值较EDA文本增强平均提升了0.02%,究其原因,EDA数据增强是基于规则对文本数据进行相关操作,虽然增强后的效果有所提升,但仍存在以下缺点:①同义词替换后的词可能与原词词向量差距不大,导致数据增强效果一般;②随机删除单词可能会删除句中的核心关键词导致与初始文本label产生偏差;③插入和替换操作在一定程度会改变句子的结构语义,在部分对句子结构有要求的任务中效果可能适得其反。而SimBERT模型通过Seq2Seq结构与相似句分类任务的联合训练方法,能够在自然语言生成任务上具有较好的表现,一定程度上提升数据的质量,但是由于生成具有一定的随机性,因而在个别模型上EDA文本增强后的效果更好。因此,由文本增强实验结果可知,文本增强技术确实能够提升分类模型的表现效果,但是总体提升效果仍然有限。

(3)RoBERTa-SimBERT模型结果分析。通过网络问政留言分类集成模型对比结果可知,经过SimBERT文本增强后的RoBERTa模型的分类效果最佳,相比于未经过文本增强的BERT-base模型和BERT+RNN模型,F1值分别提高了2.89%和3.73%。RoBERTa-SimBERT模型的详细分类结果如表6所示:

表6 RoBERTa-SimBERT模型实验结果

留言类别	查准率 P/%	查全率 R/%	F1值/%
城乡建设	91.20	93.25	92.21
劳动和社会保障	92.39	93.92	93.15
教育文体	92.86	96.59	94.69
商贸旅游	90.59	86.84	88.68
环境保护	93.12	91.19	92.15
卫生计生	93.51	89.18	91.29
交通运输	94.00	90.38	92.16
平均值(avg)	**92.52**	**91.62**	**92.05**

由表6可知,RoBERTa-SimBERT模型在网络问政留言中的分类效果总体较好,其中教育文体、劳动和社会保障以及城乡建设类别分类效果最好,F1值分别为94.69%、93.15%以及92.21%,这三类恰好也是留言数据量最多、人们最关注的问题类别。相对来说,卫生计生和商贸旅游的分类效果较差,F1值仅为91.29%和88.68%。表7是RoBERT-SimBERT模型在测试集上分类错误的典型实例,本文通过分析错误实例,试图探究不同留言类别之间分类效果产生差异的原因。

表7 RoBERTa-SimBERT模型错分类实例

留言文本	错分类别	正确类别
没有CCC认证的广告机在A市大批上市,公共安全隐患突出	城乡建设	商贸旅游
K8县天堂镇文化站蒋为将国家卫星接收器据为己有	商贸旅游	教育文体
通往K6县独坡乡坎寨村六组的路太不好走了	城乡建设	交通运输
L市禾塘村罗子坡水库食品黑作坊无证生产豆腐供入市场	卫生计生	商贸旅游
请杜绝虚假医药广告!	卫生计生	商贸旅游

从随机选择的错分类实例中可知,容易产生歧义从而影响分类器分类结果的类别主要集中在商贸旅游与卫生计生两大类,如前表6中所示,商贸旅游类的F1值仅为88.68%,而卫生计生类虽然F1值超过了91%,但是其查全率仅为89.18%。由此可知,大部分错分类实例中,将正确类别为商贸旅游的类别错分成了其他类别,导致商贸旅游类的整体分类效果较差,而被错分的类别中,错分成卫生计生类的留言量也较多,从而导致了卫生计生类的查全率较低。

通过观察错分类实例的具体留言内容,能够从定性的角度分析分类结果产生差异的原因。一方面,应该分类为商贸旅游却被错分的例子往往更加偏向"商贸"而非"旅游",这两方面确实存在理解层面的一定差异,偏向旅游主题留言更容易被正确分类,而偏向商贸主题的留言容易与其他类别,尤其是城乡建设类别混淆。另一方面,错分为卫生计生类别的留言内容中,有很多都是生产商、商铺的卫生问题,这就容易与商贸旅游的商贸主题混淆,从而影响分类的结果。

诸如上述的错分例子还有很多,分类器的错分与留言文本内容的表达息息相关,确实有部分存在歧义或者多类别的留言,这种留言影响了模型的分类效果。针对上述问题,通过文本增强对数据进行质量和数量的提升会对结果带来一定的改善,但也只能尽量维持在较佳的分类水平,想要进一步提升模型的分类效果还需从原始数据源的优化上入手。政府可以通过在其网络问政平台上对留言增加细粒度的填写限制,使得留言的内容更加详细、规范,进而提高留言分类的效果,提升效率。

5 结论

网络问政平台的兴起给了人民群众表达自身意见的渠道,通过对留言进行有效的分类可以方便政府部

门更好地把握民意，更好地自省提升。为提升问政留言分类的准确性以及提高模型端到端的部署效率，本文将预训练语言模型与文本增强技术相结合，经对比实验提出了基于 RoBERT 与 SimBERT 文本增强技术的政府网络问政留言分类模型。

传统基于预训练词向量文本特征抽取模型无法很好地处理文本“一词多义”的问题，BERT 系列预训练语言模型以其双向 Transformer 网络结构及多头注意力机制成功解决了此难题。同时，利用预训练语言模型可以有效实现端到端的模型部署，能够根据数据进行微调后将模型有效运用到多个下游任务中，基于字符级别的文本向量化方式也更加适用于中文文本表示。

此外，为有效解决问政留言领域数据质量低下的问题，本文利用了 EDA 以及 SimBERT 技术进行了文本增强来缓解训练数据不足以及数据质量问题。研究结果证明了基于 SimBERT 所生成的文本质量比基于 EDA 生成的文本质量有所提升，同时也能够解决文本分类数据增强领域的标签预测问题，为其他相关的文本增强问题提供了借鉴和思路。

与此同时，在对比实验中，将当下比较热门的基于 BERT 改进的预训练语言模型，如 BERT + RCNN、ALBERT、RoBERTa 等运用到问政留言分类问题中，并深入浅出地介绍了不同模型的优劣势及其适用范围，尝试从模型自身结构和预训练方式原理角度解释各模型之间效果产生差异的原因。BERT 系列模型原始训练语料覆盖全行业，训练质量高，可以为其他领域的文本分类问题提供具有较强借鉴及复用价值的方案和模型。

本文研究存在一定的不足之处，数据集的选择方面覆盖范围较小，主要针对的是问政留言多分类问题，之后研究可以运用不同领域的数据集进行对比以探究通用性更强的方案及模型。与此同时，可以运用相关领域语料对预训练语言模型进行微调，增加其在特定领域的文本表示与判别能力。模型结构方面，后续的研究可以尝试在 ALBERT、RoBERTa 等优秀预训练语言模型后接其他网络结构，以期取得更佳效果。

参考文献：

[1] 徐晓雯，曹守新. 网络问政对公共政策制定的影响——基于 SWOT 分析方法[J]. 山东社会科学，2015(6)：179 – 183.

[2] 马思丹，刘东苏. 基于加权 Word2vec 的文本分类方法研究[J]. 情报科学，2019，37(11)：38 – 42.

[3] 程婧，刘娜娜，闵可锐，等. 一种低频词词向量优化方法及在短文本分类中的应用[J]. 计算机科学，2020(8)：255 – 260.

[4] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the conference of the north american chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227 – 2237.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171 – 4186.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceeding of advances in neural information processing systems. California: MIT Press, 2017: 6000 – 6010.

[7] 陈燕方. 基于 DDAG-SVM 的在线商品评论可信度分类模型[J]. 情报理论与实践，2017，40(7)：132 – 137.

[8] 余本功，曹雨蒙，陈杨楠，等. 基于 nLD-SVM-RF 的短文本分类研究[J]. 数据分析与知识发现，2020，4(1)：111 – 120.

[9] 韩栋，王春华，肖敏. 基于句子级学习改进 CNN 的短文本分类方法[J]. 计算机工程与设计，2019，40(1)：256 – 260.

[10] 杨云龙，孙建强，宋国超. 基于 GRU 和胶囊特征融合的文本情感分析[J/OL]. [2021 – 02 – 10]. http://kns. cnki. net/kcms/detail/51. 1307. TP. 20200429. 1704. 010. html.

[11] 赵旸，张智雄，刘欢，等. 基于 BERT 模型的中文医学文献分类研究[J]. 数据分析与知识发现，2020(8)：41 – 49.

[12] 吴俊，程垚，郝瀚，等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究[J]. 情报学报，2020，39(4)：409 – 418.

[13] 廖胜兰，吉建民，俞畅，等. 基于 BERT 模型与知识蒸馏的意图分类方法[J/OL]. [2021 – 02 – 10]. https://doi. org/10. 19678/j. issn. 1000-3428. 0057416.

[14] JASON W, KAI Z. Eda: easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceeding of the 2019 conference on empirical methods in natural language processing. Hong Kong: ACL, 2019.

[15] 俞畅，欧阳昱，张波，等. 基于对抗式生成网络的电力用户意图文本生成[J]. 信息技术与网络安全，2019，38(11)：67 – 72.

[16] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J/OL]. [2021 – 02 – 10]. https://arxiv. org/abs/1907. 11692.

[17] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[C]//Proceedings of the international conference on learning representations. Ethiopia: ICLR, 2020.

[18] 苏剑林. 鱼与熊掌兼得：融合检索和生成的 SimBERT 模型[EB/OL]. [2020 – 05 – 18]. https://kexue. fm/archives/7427.

[19] LIU P F, QIU X P, HUANG X J. Recurrent neural for text classification with multi-task learning[C]//Proceeding of the international joint conference on artificial intelligence. New York: IJCAI, 2016.

[20] KIM Y. Convolutional neural networks for sentence classification

[C]//Proceeding of the 2014 conference on empirical methods in natural language processing. Doha: ACM, 2014.

[21] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceeding of the 29th national conference on artificial intelligence. Austin: AAAI, 2015.

[22] YANG Z L, DAI Z H, YANG Y Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceeding of the 33rd conference on neural information processing systems. Vancouver: MIT Press, 2019.

[23] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J/OL]. [2021 - 02 - 10]. https://arxiv.org/abs/1910.01108v4.

[24] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding[J/OL]. [2021 - 02 - 10]. https://arxiv.org/abs/1909.10351v3.

[25] BAO H B, DONG L, WEI F R, et al. UniLMv2: pseudo-masked language models for unified language model pre-trainning[J/OL]. [2021 - 02 - 10]. https://arxiv.org/abs/2002.12804.

作者贡献说明：

施国良：提出研究思路，修改研究方案及修订论文、定稿；

陈宇奇：设计研究方案，处理数据，构建模型及撰写论文。

A Comparative Study on the Integration of Text Enhanced and Pre-trained Language Models in the Classification of Internet Political Messages

Shi Guoliang　Chen Yuqi

Business School, Hohai University, Nanjing 211100

Abstract: [**Purpose/significance**] Government network platform for political inquiry is one of the important ways for rulers to know public opinions. In order to improve the accuracy of the classification of political inquiry messages and to deal with the problems such as poor quality and small quantity of message data, the classification effects of various BERT improved models combined with text enhancement technology and the reasons for their differences were explored. [**Method/process**] Design the network political inquiry message classification integrated comparison model, the EDA (Easier Data Augment) technology and SimBERT text Augment technology were used for comparison experiment in the aspect of text augmentation, and various pre-training language models (such as ALBERT and RoBERTa) based on BERT improvement were used for comparison experiment in the aspect of text classification model. [**Result/conclusion**] The experimental results showed that the text classification model based on RoBERTa and SimBERT text enhancement had the best effect, and the F1 value on the test set was as high as 92.05%, 2.89% higher than that of the Bert-Base model without text enhancement. At the same time, F1 value after SimBERT text enhancement was 0.61% higher than that before no enhancement. The experiment proved that text enhancement model based on RoBERTa and SimBERT can effectively improve the classification effect of multiple categories of text classification problems, and has strong referability in solving similar problems.

Keywords: political platform　text classification　text enhancement　BERT model